A reply to Rose, Livengood, Sytsma, and Machery

Chandra Sekhar Sripada[1,2], Richard Gonzalez[3,4,5], Daniel Kessler[2], Eric Laber[6], Sara Konrath[7,8], Vijay Nair[4]

[1]*Department of Philosophy, University of Michigan, Ann Arbor*
[2]*Department of Psychiatry, University of Michigan, Ann Arbor*
[3]*Department of Psychology, University of Michigan, Ann Arbor*
[4]*Department of Statistics, University of Michigan, Ann Arbor*
[5]*Department of Marketing, University of Michigan, Ann Arbor*
[6]*Department of Statistics, North Carolina State University, Raleigh*
[7]*Institute for Social Research, University of Michigan, Ann Arbor*
[8]*Department of Psychiatry, University of Rochester, Rochester, New York*

Draft Manuscript

In our paper 'Telling more than we can know about intentional action', we pursue a standard approach in the structural equation modeling field. We propose a model based on strong theoretical hypotheses and test the model using statistical tests that, in addition to being widely used in the field, have strong theoretical justification and demonstrated reliability. In their paper 'Deep Trouble for the Deep Self', Rose et al criticize our approach. In their criticism, they deploy three statistical tests, and these tests form the basis of the three main sections of their paper (i.e., sections 4, 5, and 6, respectively). The three statistical tests are:

1. Bayes Information criteria (BIC)-difference test comparing the Sripada and Konrath (S&K) theory-derived model with a model obtained through global machine learning search
2. Chi-square likelihood ratio tests (CSLRT), as well as other tests of global fit, applied separately to individual parts of the S&K model
3. Test of conditional dependence between two variables that collide on a third variable

These three tests fall well outside established practice in SEM, and appear to be novel to Rose et al.[1] In this Reply, we show that these tests devised and put forward by Rose et al contain serious errors, and the tests are in fact statistically invalid.

[1] Indeed, to the best of our knowledge, these three statistical tests lack any significant precedent in the SEM literature. Despite a fairly exhaustive search using PsychInfo and Google Scholar, we were unable to find any papers that use any of the three statistical tests listed above for the purposes of model confirmation/rejection. The CSLRT clearly has precedent in SEM, but only as a test of *overall* model fit. But after applying this, as well as other tests of overall fit, to the S&K model (and finding the S&K model fits by these tests), Rose et al then break the S&K model into parts and apply the CSLRT, as well as other tests of global fit, separately to the parts. Applying tests of global fit to parts of models is not an accepted practice. We discuss the reason why this procedure is questionable, as well as other serious problems with the authors' use of the CSLRT, in section 2 below.

1. BIC-difference test

Rose et al perform a search using TETRAD IV (Spites, Glymour, and Sheines, http://www.phil.cmu.edu/projects/tetrad/tetrad4.html), a machine learning technique for the discovery of causal structure[2], and find a model that is supposed to have better fit than the Sripada and Konrath model (Figure 1). The differences in fit are modest; the TETRAD-outputted model has a BIC that is only 3.39 points better than the S&K model.[3] However, Rose et al believe this difference in BIC (and closely related fit statistics) licenses the rejection of the S&K model in favor of the TETRAD-outputted model. They write:

> When faced with non-hierarchical models, one typical practice is to choose the model with the best AIC or BIC score (Kaplan 2009; Klein 1998; Loehlin 2004; Rafferty 1995; Raykov and Marcoulides 2000; Rust et al. 1995; Schreiber et al. 2006). Following this practice, we would pick the Tetrad models over the S&K models… [W]e conclude that the data undermine the Deep Self Concordance Account, as it is currently formulated (pg. 9-10).

The inference Rose et al make is not statistically valid. A search-outputted model will inevitably be to some degree overfit to the data and thus its BIC will be spuriously elevated. When undertaking a BIC comparison, one must adjust the BIC of the search-outputted model downwards to account for overfitting in order to make a fair comparison with an alternative model. A bit later, we will discuss how one might get a sense of the magnitude of the adjustment required. But in the absence of such an adjustment, Rose et al.'s use of unadjusted BIC differences for the purposes of model rejection is not legitimate.[4]

It is useful to explain the issue further in intuitive terms. Data obtained from an experiment can be seen as arising from a combination of two sources, the true underlying causal processes operative in nature and random noise. Machine learning procedures such as those implemented in TETRAD use the data itself to arrive at a model, and thus will inevitably find a model that accommodates not only the true causal variation but also the noise-related variation, and as a result, the TETRAD-outputted model will have a spuriously elevated fit. Because of this problem of overfitting, many machine learning applications make quantitative

---

[2] We note at the outset that our criticism is not directed at the TETRAD approach to causal discovery. We view TETRAD as an excellent search tool that is underused in the social sciences. The problem is that Rose et al attempt to turn goodness of fit measures associated with TETRAD-outputted models into a statistical test for rejection of competing theory-based models. It is *specifically* this statistical test devised by Rose et al that is not legitimate.

[3] Most researchers would consider this to be too small a BIC difference to warrant serious attention. Indeed, this is the view of one of the authors of TETRAD himself (Clark Glymour, personal communication). So our present argument should be seen as a hypothetical one: *Were it the case that the BIC difference were sizable*, even then the Rose et al argument is not statistically legitimate.

[4] The authors Rose et al cite in the preceding quoted paragraph are all using the BIC-difference test to compare two theory-derived models. Rose et al are thus taking these authors very much out of context when they cite these authors as supporting use of the BIC-difference test to compare a theory-derived model with a *machine learning global search-outputted model*.

interpretations of the fit of search-outputted models only after some validation procedure, for example training using only a subset of the data, and testing on the remaining data (i.e., cross-validation). Rose et al fail to do this, and, remarkably, do not discuss the issue of overfitting at all.

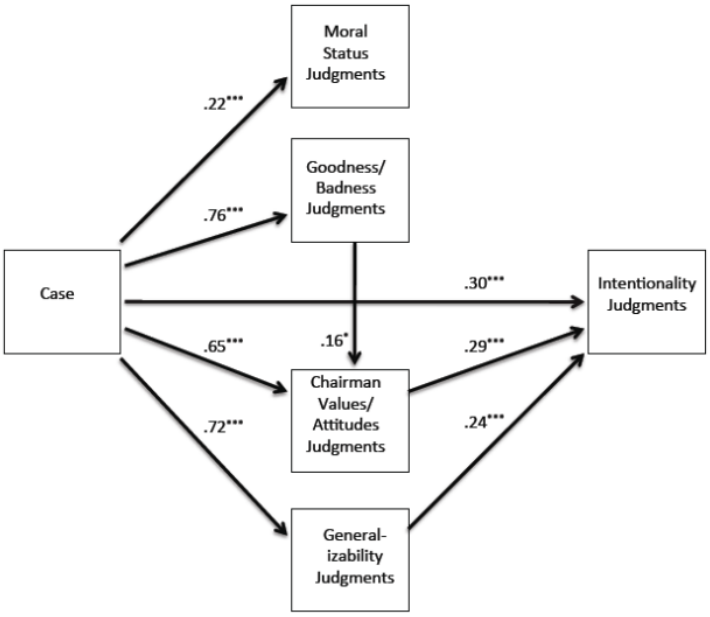

Figure 1: The Sripada and Konrath (S&K) model

How large is the spurious improvement in the BIC that arises due to overfitting? We know of no analytic methods by which its magnitude might be calculated. However, Clark Glymour, one of the authors of TETRAD, has suggested a simulation can be run to get a better sense of its size (personal communication). Based on Glymour's suggestions, we ran a Monte Carlo simulation that proceeded as follows:

1. Assume the S&K model from Figure 1 is the true model.
2. Generate data under this model (240 cases as in the original S&K study), using sample variances of the variables from the S&K study to establish the relevant distributions.
3. Conduct a TETRAD Greedy Equivalent Search (GES) on the simulated data.
4. Apply the S&K model and the TETRAD-outputted model to the simulated data, and compare the BICs of these models. In particular, record the improvement in BIC between the TETRAD-outputted model and the S&K model.
5. Repeat these steps to generate a total of 1,000 simulated data sets

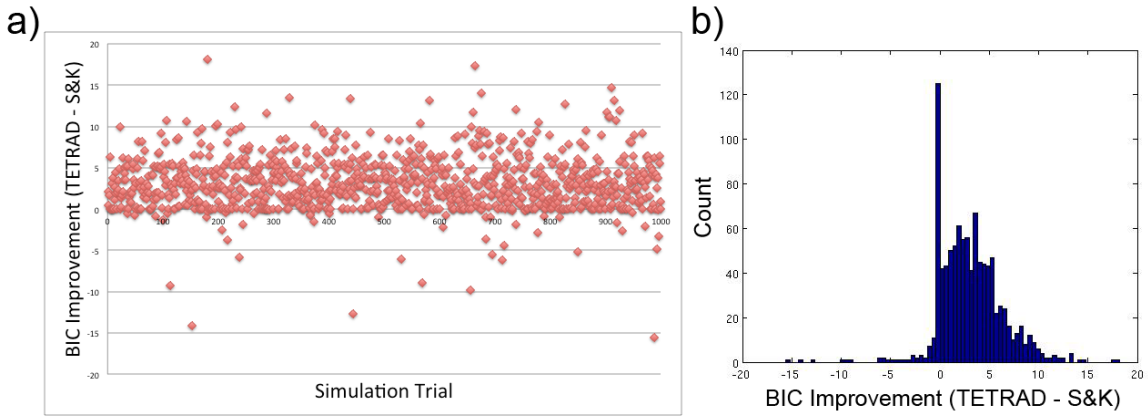

Figure 2: Spurious improvement in the Bayes Information Criterion (BIC) in a simulation of 1000 sets of data. Data were simulated under the Sripada and Konrath (S&K) model from Figure 1 (240 observations for each set of data). A TETRAD IV search using the Greedy Equivalent Search (GES) algorithm was performed on the simulated data sets. The BIC scores from the model outputted from the TETRAD search were compared to the BIC of the S&K model. a) Trial by trial values for the BIC improvement; b) the distribution of BIC improvement values.

In Figure 2a, we show the differences in BIC values for each run of the simulation.[5] The y-axis shows the BIC for the TETRAD-outputted model fitted to the data set minus the BIC for the true (S&K) model fitted to the same data. Figure 2b is a histogram of these differences, with the BIC improvement now plotted on the x-axis. Note that most of these differences are positive (indicating that the fitted TETRAD model is declared to be better than the true S&K model, i.e. the model that generated the simulated data). Moreover, these 'winning' models found by Tetrad exhibited substantial variability in their causal graphs across simulation trials, highlighting that these models capitalize on chance variation present in individual samples (i.e., these models are overfit to the data). The mean BIC difference was 3.10, and the 50th and 95th percentiles were 2.78 and 8.92. Recall that the BIC difference observed between the TETRAD-outputted model identified by Rose et al. and the S&K model was 3.39, quite close to the center of the distribution in Figure 2b. Further, we can calculate from Figure 2b that, if one followed Rose et al.'s counsel to always prefer the model with the lower BIC, one would reject the true model 83% of the time. So we see that Rose et al.'s unqualified use of BIC differences for the purposes of model rejection is not valid and will lead to the wrong conclusions.

[5] A single outlier BIC value more than 10 standard deviations from the mean was excluded from the analysis. This appears to represent a case in which the TETRAD search procedure was stuck at a local optimum.

It is worth re-emphasizing that we are *not* criticizing the TETRAD machine learning approach to causal discovery more generally. Our arguments and simulations are directed at criticizing, quite specifically, Rose et al's attempt to use TETRAD as a procedure for falsifying alternative theory-based models. It is noteworthy that the authors of TETRAD avoid claims that the output of a TETRAD search can be directly used to falsify alternative theory-based models.[6] They instead focus on using TETRAD as a tool that can reliably identify causal structure, especially in large sample sizes, and that can pry theorists away from excessive reliance on possibly shaky a priori assumptions. If Rose et al had used TETRAD to show that there is an alternative to the S&K model that warrants serious consideration, then we would be the first to recognize this as an important finding. We view TERTAD as an outstanding search and modeling tool that is vastly underutilized by the social science community. Our criticism, therefore, is directed at Rose et al's invalid inference from the output of a TETRAD search, and not the TETRAD approach itself.

2. Chi-square likelihood ratio tests (CSLRT), as well as other tests of global fit, applied separately to individual parts of the S&K model

The argument presented in Section 1 already invalidates the analysis conducted by Rose et al. But to be complete, we also briefly discuss the two other statistical tests used by Rose et al.

Rose et al apply tests of global fit including the chi-square likelihood ratio test (CSLRT) to the S&K model and find the model has excellent fit to the data by these tests. They then divide the S&K model into two parts, which they call the Negative Sub-Model and the Positive Sub-Model. They then re-apply the tests of global fit, including the CSLRT, to each part of the model. A good fitting model should yield a non-significant value for the CSLRT. They find that for the Positive Sub-Model, the CSLRT is significant at $p = 0.04$. They then claim that the Deep Self Concordance Account is undermined by the data (pg. 11).

There are multiple serious problems here. First, the legitimacy of breaking models into parts and applying tests of global fit on each part is highly questionable. This falls well outside SEM practice and Rose et al provide no justification for why this procedure was used in lieu of the established approaches[7], and whether this

[6] See for example Sheines, Sprites et al (1998, pg. 102 and pg. 105), and Chapter 10 of Sprites, Glymour, et al (2000).

[7] Rose et al say their purpose is applying tests of global fit to individual parts of models is that they are trying to identify local model misspecification, and in particular whether the parts of the S&K model associated with different sub-hypotheses exhibit good fit. But it is unclear why Rose et al do not use, or even *mention*, the standard and recommended practice in the SEM field to investigate local model misspecification – examination of the residuals of the covariance matrices. Inspection of residuals revealed no correlation residuals greater than 0.1 (Kline 2005) and no standardized covariance residuals greater than 2.58 (Joreskog and Sorbom 1993), the recommended cut-offs in the field. This means the S&K model fits the data according to standard criteria *without evidence of local model misspecification*, which is precisely what Rose et al say they were trying to ascertain.

procedure is statistically valid and reliable. Indeed, the natural culmination of such an approach is that one should apply tests of global model fit separately to each individual path in an SEM model – a patently absurd result.[8]

Second, the CSLRT has been roundly criticized in the SEM field, and the clear norm in the field is that one should *not* use the CSLRT as a sole test of model confirmation/rejection. In a widely respected review of model fit in SEM, Kenneth Bollen and Scott Long write:

> A second point of consensus is that the chi-square test statistic should not be used as the sole basis for determining model fit. Several reasons support this belief. First, the null hypothesis underlying the test statistic is overly rigid… A second reason is that the chi-square test statistic as usually applied ignores the statistical power of the test. Third, failure of the variables to satisfy the distributional assumptions of the test can lead to rejection of correct models…(Bollen and Long 1993, pg. 6).

There are multiple problems with the CSLRT[9], but we focus on Bollen and Scott's third point: the CSLRT is highly sensitive to deviations from *multivariate normality* (in particular, multivariate kurtotic distributions inflate type 2 errors, i.e., rejection of true models). This is germane to the S&K study because, given that the data for this study was collected on a 7-point Likert scale, deviations from multivariate normality are practically unavoidable. Interestingly, robust procedures are available to correct for the effect of multivariate non-normality. We used a Satorra-Bentler correction implemented in the EQS SEM package (Multivariate Software Inc, Encino CA). Rose et al should have applied a correction of this sort before relying on the

---

[8] At least one obvious problem with evaluating the overall fit of a model and then subsequently separately evaluating the fit of various parts of the model is that, like any sub-group analysis, this strategy requires the proliferation of post-hoc statistical tests, generating a problem of multiple comparisons. To address this, Rose et al should have applied a Bonferroni correction. When this is done, the Positive Sub-Model is in fact *supported* by the CSLRT.

[9] Regarding Bollen and Scott's first point, the CSLRT tests the null hypothesis that a model fits the data *perfectly* (i.e., the test reports the likelihood of the following hypothesis: 'There is *zero* deviation between the implied and sample covariance matrices'). This is not the null hypothesis most researchers are interested in testing. It is precisely for this reason that tests of approximate fit have been developed, and it is noteworthy that the S&K Positive Sub-model scores well on according to these alternative tests. Another problem not mentioned by Bollen and Long is that the CSLRT is highly sensitive to the magnitude of the paths in the model. This is germane to the S&K model because the path coefficients that are relevant to computation of the test statistic are quite large (0.6-0.8), inflating type 2 errors (rejection of correct models) dramatically. Because of these problems, in actual practice, most SEM researchers ignore the CSLRT as a test of model fit. For example, a survey of SEM studies in the journal *Personality and Individual Differences* found that 25 of 28 studies reported a model that fails by the CSLRT (Markland 2007). In one of the most widely used textbooks for SEM (Byrne 2010; see also Byrne 2010), the author tests a model with data from 260 subjects (a sample size quite similar to the S&K study) and obtains a range of fit statistics, as well as other indices of model fit such as residual covariances and modification indices. She finds that the CSLRT is *highly* significant ($p<0.001$). However, she concludes that given the dubious value of the CSLRT, the plausibility and significance of the paths, good fit by other indices, and given that modification tests are not significant, the model should be accepted.

CSLRT as a strict test of model rejection. After correction, we found the CSLRT applied to the Positive Sub-Model was *not* statistically significant [S-B $X^2(1, N = 240) = 2.68$, $p > 0.1$], that is the CSLRT *supports* the S&K positive sub-model (even without the multiple comparison correction discussed in footnote 8).

We can thus sum our response to Rose et al's application of the CSLRT and other tests of global fit to parts of the S&K model this way. It is a seriously questionable practice to apply tests of overall model fit to parts of a model. But even if one *does* attempt to apply a test of overall model fit in this way, the CSLRT is *not* recommended as a sole test of model rejection. But even if one *does* insist on using this test as the sole criterion of model rejection, Rose et al did not execute the test correctly (they fail to correct for multiple comparisons and for multivariate non-normality). When the CSLRT is executed correctly, the S&K Positive Sub-Model clearly passes this test.

3. Test of conditional dependence between two variables that collide on a third variable

Rose et al claim that if we assume certain conditions hold (in particular, the Markov and Faithfulness conditions; details about the meaning of these conditions are not important for the present point) then we should observe certain patterns of graph theoretic implied patterns of conditional dependence. In particular, they note the following:

(D) If the S&K model is the true causal model, then Chairman Values/Attitudes and Generalizability should be dependent, conditional on both Intentionality Judgments and Case. [10]

D is certainly true. But a problem arises because Rose et al erroneously believe that the truth of D can be leveraged into a statistical test for model confirmation/rejection. A red flag about this proposed statistical test arises immediately because: 1) it is not an established practice in SEM to use tests of conditional dependence of this sort for model confirmation/rejection (indeed, we are aware of no precedent for this at all); 2) no references are offered to papers that discuss the test and provide validation; and 3) Rose et al themselves provide no evidence that this novel test they have put forward is reliable.

As it turns out, the test that Rose et al put forward is clearly unreliable. The problem is this: If the S&K model is true, while one should expect a dependence between Chairman Values/Attitudes and Generalizability (conditional on both Intentionality

[10] Rose et al make the additional claim that Chairman Values/Attitudes and Generalizability should be independent conditional on Case. Note that mathematically this is identical to the CSLRT we discussed in Section 2 (notice the *p*-values reported by Rose et al for these tests are the same), though it appears that Rose et al do not realize this equivalence. Our response to their argument regarding the CSLRT thus applies to the test of conditional independence as well.

Judgments and Case), *the magnitude of the dependence will be extremely small*. Given the implied covariance matrix of the S&K model, it is calculated to be 0.057 in standardized units, and a dependence of this size cannot be reliably detected (that is, at the relevant sample sizes, power is unacceptably low to detect the effect).

We can again use Monte Carlo simulation to illustrate the lack of power, and thus unreliability, of the test devised by Rose et al. The test proposed by Rose et al rejects the S&K model if the model-implied dependence between Chairman Values/Attitudes and Generalizability (conditional on both Intentionality Judgments and Case) *is not* detected. A simulation of 1000 data sets generated under the S&K model finds that in only 21% of the cases, a statistically significant dependence was detected between Chairman Values/Attitudes and Generalizability, conditional on Intentionality Judgments and Case (setting alpha at the standard 0.05). The remaining 79% of the time, the small dependence between these two variables fails to be detected*, even though this dependence is present in the model from which this data is generated*. This demonstrates that the statistical test by Rose et al is grossly underpowered rendering it unacceptable as a test for model rejection. Even when the S&K model is known to be true, the statistical test devised by Rose et al detects the true model only 21% of time, while there is a 79% probability of making a Type II error (i.e., rejection of the true model).

Summary

We have reviewed the three statistical tests used by Rose et al. It is notable that these tests fall well outside established practice in SEM, and appear to be novel to Rose et al. We have demonstrated all three tests are statistically invalid.

References

Bollen, K. A. and J. S. Long (1993). Introduction. Testing Structural Equation Models. K. A. Bollen and J. S. Long. New Bury Park, CA, SAGE Publications.

Byrne, B. (2010). Structural Equation Modeling with AMOS. New York, Routledge.

Byrne, B. (2010). Structural Equation Modeling with EQS. New York, Routledge.

Joreskog, K. G. and D. Sorbom (1993). LISREL 8: structural equation modeling with the SIMPLIS command language, Scientific Software International.

Kline, R. B. (2005). Principles and practice of structural equation modeling. New York, Guilford Press.

Markland, D. (2007). "The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modelling." Personality and Individual Differences **42**(5): 851-858.

Scheines, R., P. Spirtes, et al. (1998). "The TETRAD project: Constraint based aids to causal model specification." Multivariate Behavioral Research **33**(1): 65-117.
Sprites, P., C. Glymour, et al. (2000). Causation, Prediction, Search. Cambridge, MA, MIT Press.